which hold for any $(\overline{\pi}, \overline{\mu})$. This suggests the estimator

$$\widehat{\psi}_{dr} = \mathbb{P}_n \left[\left\{ \frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)} \right\} \Big\{ Y - \widehat{\mu}_A(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \widehat{\mu}_0(X) \Big\} \right] \tag{4.4}$$

which was also used in the previous chapter with experiments, except now the propensity score $\pi(x)$ depends on covariates and is unknown so needs to be estimated. The doubly robust estimator is somewhat less intuitive than the other two options, but it can be viewed as correcting leftover smoothing bias of a regression of inverse-probability-weighted estimator, or augmenting an inverse-probability-weighted estimator with regression predictions to increase efficiency. We will see in later chapters that its precise form comes from a bias correction based on a distributional Taylor expansion of the average treatment effect functional.

4.4.1 Discrete Covariates

For some intuition we will first consider the simplest case, where the covariates X are discrete and low-dimensional, i.e., $X \in \{1, ..., d\}$ with d fixed. We will see that in this setup, when one uses the empirical distribution to estimate the "nuisance functions" π and μ_a, then all three of the previously mentioned estimators coincide in that they are numerically equivalent. (Later we will show that they are asymptotically efficient in a local minimax sense). This numerical equivalence does not occur when the covariates have some continuous components and modeling or smoothing is used to construct the $\widehat{\pi}$ and $\widehat{\mu}_a$ estimates. Intuitively, the reason why all three estimators are numerically equivalent is because, when the covariates are discrete, there is no smoothness or additional structure to exploit, so each estimator is making full equivalent use of the data. Another way to think about it is that, in the discrete case, the empirical measure $\mathbb{P}_n$ is an actual valid distribution (including all conditional distributions), and so the identifying expression equalities above also hold for $\mathbb{P}_n$.

Our first result shows the numerical equivalence between the regression, weighting, and doubly robust estimators.

Proposition 4.4. *Suppose $X \in \{1, ..., d\}$ is discrete and the nuisance estimators are the empirical averages*

$$\begin{aligned} \widehat{\pi}(x) &= \mathbb{P}_n(A \mid X = x) = \frac{\mathbb{P}_n\{A\mathbb{1}(X = x)\}}{\mathbb{P}_n\{\mathbb{1}(X = x)\}} \\ \widehat{\mu}_a(x) &= \mathbb{P}_n(Y \mid X = x, A = a) = \frac{\mathbb{P}_n\{Y\mathbb{1}(A = a)\mathbb{1}(X = x)\}}{\mathbb{P}_n\{\mathbb{1}(A = a)\mathbb{1}(X = x)\}} \end{aligned}$$

Then the regression, weighting, and doubly robust estimators defined in (4.2)–(4.4) are all numerically equivalent, i.e.,

$$\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}.$$

Proof. We will consider the $\psi_1 = \mathbb{E}\{\mu_1(X)\}$ term, since the logic is the same for ψ_0. To see that $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw}$ note that

$$\begin{aligned}
\widehat{\psi}_{reg} &= \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(X_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{P}_n\{YA\mathbb{1}(X = x_i)\}}{\mathbb{P}_n\{A\mathbb{1}(X = x_i)\}} \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{1}{n}\sum_j Y_j A_j \mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)\}} = \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{\frac{1}{n}\sum_i \mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_j)} = \frac{1}{n}\sum_{j=1}^{n} Y_j A_j / \widehat{\pi}(X_j) = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(X)}\right\} = \widehat{\psi}_{ipw}
\end{aligned}$$

where in the fifth equality we replace the x_i in the denominator with x_j since the numerator includes the indicator $\mathbb{1}(X_j = x_i)$.

Now to see that $\widehat{\psi}_{reg} = \widehat{\psi}_{dr}$ we will show $\mathbb{P}_n\{AY/\widehat{\pi}(X)\} = \mathbb{P}_n\{A\widehat{\mu}_1(X)/\widehat{\pi}(X)\}$, so that the correction term $\mathbb{P}_n[A\{Y - \widehat{\mu}_1(X)\}/\widehat{\pi}(X)] = 0$. Note

$$\begin{aligned}
\mathbb{P}_n\left\{\frac{A\widehat{\mu}_1(X)}{\widehat{\pi}(X)}\right\} &= \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\widehat{\pi}(X_i)} \frac{\frac{1}{n}\sum_j Y_j A_j \mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{1}{n}\sum_{i=1}^{n} \frac{A_i}{\widehat{\pi}(X_i)} \frac{\mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_i)} \\
&= \frac{1}{n}\sum_{j=1}^{n} Y_j A_j \frac{1}{\widehat{\pi}(X_j)} \frac{1}{n}\sum_{i=1}^{n} A_i \frac{\mathbb{1}(X_j = x_i)}{\frac{1}{n}\sum_k A_k \mathbb{1}(X_k = x_j)} \\
&= \frac{1}{n}\sum_{j=1}^{n} \frac{Y_j A_j}{\widehat{\pi}(X_j)} = \mathbb{P}_n\left\{\frac{AY}{\widehat{\pi}(X)}\right\}
\end{aligned}$$

where again in the third equality we replace x_i in the denominator with x_j due to the numerator indicator. This gives the result. □

Next we derive the limiting distribution of the estimator $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}$.

Theorem 4.1. *Suppose $X \in \{1, ..., d\}$ is discrete and the nuisance estimators are the empirical averages from Proposition 4.4. Assume that Y is bounded and that $\pi(x)$ and $\widehat{\pi}(x)$ are bounded away from ϵ and $1 - \epsilon$ for some $\epsilon > 0$ and all x. Then*

$$\sqrt{n}(\widehat{\psi} - \psi) \rightsquigarrow N(0, var(f))$$

for $\widehat{\psi}$ the estimators in (4.2)–(4.4) and

$$f(Z) = \mu_1(X) - \mu_0(X) + \left\{\frac{A}{\pi(X)} - \frac{1 - A}{1 - \pi(X)}\right\}\Big\{Y - \mu_A(X)\Big\}.$$

Proof. We will work with the $\widehat{\psi}_{dr}$ version of the estimator, which can be written as $\widehat{\psi}_{dr} = \mathbb{P}_n(\widehat{f})$ for

$$f(Z) = \mu_1(X) - \mu_0(X) + \left\{ \frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)} \right\} \Big\{ Y - \mu_A(X) \Big\}$$

and $\widehat{f}$ the version of f replacing (π, μ_a) with $(\widehat{\pi}, \widehat{\mu}_a)$.

Therefore by Lemma 3.1 we have the decomposition

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})f + (\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) + \mathbb{P}(\widehat{f} - f) \equiv Z^* + T_1 + T_2.$$

We will first handle the T_1 term. Note since X is discrete we can write the nuisance estimators $(\widehat{\pi}, \widehat{\mu}_a)$ as linear regression estimators based on saturated models, i.e.,

$$\widehat{\pi}(x) = \pi(x; \widehat{\alpha}) = \widehat{\alpha}^{\mathrm{T}} w$$

where $w^{\mathrm{T}} = \{\mathbb{1}(x=1), ..., \mathbb{1}(x=d-1)\} \in \{0,1\}^{d-1}$ and similarly

$$\widehat{\mu}_a(x) = \mu_a(x; \widehat{\beta}_a) = \widehat{\beta}_a^{\mathrm{T}} w.$$

This implies

$$|\widehat{f}(z) - f(z)| = |f(z; \widehat{\eta}) - f(z; \eta)| \leq C \|\widehat{\eta} - \eta\|$$

for $\eta = (\alpha, \beta_0, \beta_1)$ and $C < \infty$ some constant. Therefore f and $\widehat{f}$ belong to a Donsker class, which together with the central limit theorem and Lemma 19.24 from van der Vaart [2000] imply that $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$.

For the T_2 term, note that $f = f_1 - f_0$ for $f_a = \mu_a + \frac{\mathbb{1}(A=a)(Y-\mu_a)}{a\pi(x)+(1-a)\{1-\pi(x)\}}$. Then

$$\begin{aligned}
\mathbb{P}(\widehat{f}_1 - f_1) &= \mathbb{P}\left[\frac{A}{\widehat{\pi}(X)} \Big\{ Y - \widehat{\mu}_1(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \mu_1(X) \Big\} \right] \\
&= \mathbb{P}\left[\frac{\pi(X)}{\widehat{\pi}(X)} \Big\{ \mu_1(X) - \widehat{\mu}_1(X) \Big\} + \Big\{ \widehat{\mu}_1(X) - \mu_1(X) \Big\} \right] \\
&= \mathbb{P}\left[\frac{\pi(X) - \widehat{\pi}(X)}{\widehat{\pi}(X)} \Big\{ \mu_1(X) - \widehat{\mu}_1(X) \Big\} \right] \\
&\leq \mathbb{P}\left\{ \left| \frac{\pi(X) - \widehat{\pi}(X)}{\widehat{\pi}(X)} \right| \Big| \mu_1(X) - \widehat{\mu}_1(X) \Big| \right\} \\
&\leq \left(\frac{1}{\epsilon} \right) \mathbb{P}\left\{ \Big| \pi(X) - \widehat{\pi}(X) \Big| \Big| \mu_1(X) - \widehat{\mu}_1(X) \Big| \right\} \\
&\leq \left(\frac{1}{\epsilon} \right) \|\pi - \widehat{\pi}\| \|\mu_1 - \widehat{\mu}_1\| \\
&= O_{\mathbb{P}}(1/\sqrt{n}) O_{\mathbb{P}}(1/\sqrt{n}) = O_{\mathbb{P}}(1/n) = o_{\mathbb{P}}(1/\sqrt{n})
\end{aligned}$$

where the second and third lines used iterated expectation, the fifth used the bound on $\widehat{\pi}$, the sixth used Cauchy-Schwarz, and the last line used that $\widehat{\pi}$ and $\widehat{\mu}_a$ are root-n consistent due to the discrete (e.g., they can be represented as linear regression estimators, as mentioned above). The same exact logic follows for $\mathbb{P}(\widehat{f}_0 - f_0)$, which then yields the result since $T_1 + T_2 = o_{\mathbb{P}}(1/\sqrt{n})$. □

Theorem 4.1 shows that, when the covariates are discrete and low-dimensional, the causal effect estimators $\widehat{\psi}_{reg} = \widehat{\psi}_{ipw} = \widehat{\psi}_{dr}$ are all root-n consistent and asymptotically normal under only mild boundedness conditions. The key to proving this result was the analysis of the T_2 term; the logic used there will be repeated throughout the book going forward.

Theorem 4.1 gives confidence intervals (and thus hypothesis tests) as an immediate corollary.

Corollary 4.1. *Under the conditions of 4.1, an asymptotically valid confidence interval for the average treatment effect ψ is given by*

$$\widehat{\psi} \pm 1.96\sqrt{\widehat{var}(\widehat{f})/n}.$$

Remark 4.4. Although the regression, weighting, and doubly robust estimators are exactly equal, to construct confidence intervals we need to estimate the asymptotic variance with the empirical variance of the terms appearing in the doubly robust estimator.

In summary, when the measured covariates are sufficient to control confounding, and are discrete and low-dimensional, the choice of estimator is immaterial – regression, weighting, and doubly robust estimation are all numerically equivalent and efficient (note though that a simple difference-in-means estimator is no longer even consistent). In the next section, however, we will see that the story is much different in the more realistic scenario where the covariates are not all discrete, and so some modeling is necessary.

4.4.2 Regression & Matching

As mentioned earlier, the regression estimator (4.2) is perhaps most intuitive, since it immediately follows from plugging estimates into the identification result from Proposition 4.2. In practice, as in experiments with covariate adjustment, one could use simple parametric estimators (e.g., linear or logistic regression) or more flexible nonparametric estimators (e.g., kernel smoothing, random forests) of the regression functions μ_a.

First we will consider the case where μ_a is estimated with a finite-dimensional parametric model, i.e., it is assumed that

$$\mu_a(x) = \mu_a(x;\beta)$$

for some real-valued parameter $\beta \in \mathbb{R}^p$. A prominent example might include a logistic regression model $\mu_a(x;\beta) = \text{expit}(\beta_0 + \beta_1 a + \beta_2^{\mathrm{T}} x)$. The parameter β could be estimated via maximum likelihood or some relevant m-estimator, for example. Note that the discrete covariate setup can be viewed as a special case of this, since then for each $a = 0, 1$ the function $\mu_a(x)$ can be estimated with a saturated model with d parameters (e.g., $d-1$ level indicators and an intercept).

In fact, we have already analyzed the estimator (4.2) in the case where it is assumed that μ_a follows a parametric model, in Theorem 3.1. Recall when we analyzed the parametric plug-in estimator we did not rely on the randomization at all, so the same analysis applies here. The relevant theorem is repeated below for posterity.

Theorem 4.2. *Let $f(x) = \mu_1(x) - \mu_0(x)$ and $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$, so that $\psi = \mathbb{E}\{f(X)\}$ is the average treatment effect. Assume the parametric model*

$$\mu_a(x) = \mu_a(x;\beta_a)$$

for some $\beta = (\beta_0, \beta_1) \in \mathbb{R}^p$. Suppose the estimator $\widehat{\beta} \in \mathbb{R}^p$ solves an estimating equation so that

$$\mathbb{P}_n\{m(Z;\widehat{\beta})\} = 0.$$

Assume $m(z;\beta) \in \mathbb{R}^p$ is Lipschitz in β, and that $\mathbb{E}\{m(z;\beta)\}$ is differentiable at the true β satisfying $\mathbb{E}\{m(Z;\beta)\} = 0$ with nonsingular derivative matrix. Then

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z;\beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z;\beta) = f(x;\beta) + \frac{\partial \mathbb{E}\{f(X;\beta)\}}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(z;\beta)$$

and so is root-n consistent and asymptotically normal.

Theorem 3.1 shows that if a correct parametric model is available for the regression functions μ_a, and under some regularity conditions (essentially smoothness of the model), then the resulting plug-in regression estimator is root-n consistent and asymptotically normal. Confidence intervals can be constructed using an estimate of the closed-form asymptotic variance $\widehat{\text{var}}\{\widehat{g}(Z;\widehat{\beta})\}$, or via the bootstrap which is typically easier. Bootstrap estimates would be computed in the usual way: take a bootstrap sample (i.e., sample n observations with replacement), re-estimate the regression functions μ_a, and construct the plug-in estimator (and then repeat many times).

In practice, there is often not enough detailed background knowledge to justify a particular parametric model, especially in the presence of continuous covariates. This motivates the use of a nonparametric version of the regression estimator (4.2). However, in the nonparametric case, the analysis from Theorem 3.1 no longer applies, since μ_a is no longer assumed to be indexed by a finite-dimensional parameter. This means that the term $T_2 = \mathbb{P}(\widehat{f} - f)$ in the proof cannot be analyzed by simply differentiating with respect to β.

In fact, our discussion from Section 3.3.3 concerning nonparametric covariate adjustment in experiments applies as well, since as in the parametric case we did not rely on randomization there. To reiterate: when the regression functions μ_a are estimated nonparametrically, the term $T_2 = \mathbb{P}(\widehat{f} - f)$ cannot in general be expected to be $O_{\mathbb{P}}(1/\sqrt{n})$, so that the regression estimator would typically inherit slow convergence rates from estimating μ_a flexibly. This means that a nonparametric version of the regression estimator (4.2) has advantages over a parametric version because it will generally be consistent under weaker conditions, and thus more robust; however it has a potential disadvantage in that it will be less efficient, compared to if the parametric model is correctly specified. Further, it is not only slow rates that complicate the use of this estimator: for general nonparametric estimators it will often not have a tractable limiting distribution, and even when it does it would typically not be correctly centered (without undersmoothing), and so inference would be complicated at best.

Importantly, however, there are some exceptions to this story. Namely, if one specifies particular nonparametric estimators $\widehat{\mu}_a$ (e.g., based on kernels, splines, or nearest-neighbor regression), undersmooths (so that smaller bias of $\widehat{\mu}_a$ is traded off for larger variance, rather than the usual balancing), and makes some particular structural assumptions (e.g., that the regression functions are Hölder smooth), then nonparametric regression estimators can be root-n consistent (and sometimes asymptotically normal).

We will first discuss such an analysis of matching estimators, and then briefly mention series/spline estimators.

Matching

Matching is often treated as an entirely different adjustment approach, but it can be viewed as a particular nonparametric regression-based estimator, for example using k-nearest neighbor regression to estimate μ_a. Following Abadie and Imbens [2006], one simple version of matching uses $\widehat{\psi}_{reg}$ from (4.2) with

$$\widehat{\mu}_{A_i}(x_i) = Y_i \quad , \quad \widehat{\mu}_{1-A_i}(x_i) = \frac{1}{K} \sum_{j \in \mathcal{J}_K(i)} Y_j$$

where $\mathcal{J}_K(i)$ is the set of K indices with the treatment opposite of a_i, who are closest in terms of covariates, i.e., $\mathcal{J}_K(i) = \{j_1(i), ..., j_K(i)\}$ for

$$j_k(i) = \left\{ j : \sum_{\ell : A_\ell \neq A_i} \mathbb{1}\{\|X_\ell - X_i\|_2 \leq \|X_j - X_i\|_2\} = k \right\}$$

the index of the k^{th} closest match, for $\| \cdot \|_2$ the Euclidean norm. This corresponds to matching with replacement, since the same indices can reappear in $\mathcal{J}_K(i)$ for different subjects i. Of course, other schemes can also be used, e.g., based on matching without replacement, or via other distance metrics, or using a variable number of neighbors k, etc.

Abadie and Imbens [2006] show that, under usual iid sampling setup and consistency/exchangeability/positivity assumptions, and if

- the covariates $X \in \mathbb{R}^d$ are continuous, have compact and convex support, and density bounded away from zero, and if
- $\mu_a(x)$ is Lipschitz in x for $a \in \{0, 1\}$,

then for fixed K the matching estimator has (conditional) bias

$$\mathbb{P}_n \left((2A-1) \Big[\mu_{1-A}(X) - \mathbb{E}\{\widehat{\mu}_{1-A}(X) \mid X, A\} \Big] \right) = O_{\mathbb{P}}(1/n^{1/d})$$

whose expected value is not of smaller order than $O(1/n^{2/d})$. This means that the estimator $\widehat{\psi}_{reg}$ when based on K-nearest neighbor matching is

- root-n consistent and asymptotically normal, if $d = 1$;
- root-n consistent but not necessarily asymptotically normal, if $d = 2$;
- not root-n consistent or asymptotically normal, if $d > 2$.

This illustrates the kind of result we discussed above heuristically: in general nonparametric versions of the regression estimator $\widehat{\psi}_{reg}$ are not expected to be root-n consistent and asymptotically normal, but some exceptions can occur, e.g., for particular K-nearest neighbor estimators, under smoothness conditions and/or low dimensions.

There are other examples of nonparametric estimators being root-n consistent with undersmoothing. Hahn [1998] has a nice paper that explores efficiency bounds for average treatment effects, and considers undersmoothed regression estimators based on series/spline regression. He shows that: if one uses particular orthonormal bases, and carefully tuned number of series terms, and if the propensity score and regression functions are infinitely differentiable, then a regression-type estimator is root-n consistent and asymptotically normal. Some caveats of these kinds of analyses are that in practice:

- it is difficult to construct bases satisfying the required regularity conditions;
- it can be even more difficult to pick the right number of basis terms (i.e., it requires undersmoothing – suboptimal tuning by trading less bias for more variance).

4.4.3 Weighting

The inverse-probability-weighted estimator (4.3) can be motivated from importance sampling or representativeness arguments, and depends on an estimate of the unknown propensity score $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$. In fact, its analysis is essentially the same as that of the regression-based estimator; the main difference is we define

$$\widehat{f} = \left\{ \frac{A}{\widehat{\pi}(X)} - \frac{1-A}{1-\widehat{\pi}(X)} \right\} Y$$

instead of $\widehat{f} = \widehat{\mu}_1 - \widehat{\mu}_0$.

First we will consider the case where the propensity score π is estimated with a parametric model, i.e.,

$$\widehat{\pi}(x) = \pi(x; \widehat{\beta}).$$

For example with logistic regression we might have $\pi(x;\beta) = \text{expit}(\beta_0 + \beta_1^{\mathrm{T}} x)$. As with the regression estimator $\widehat{\psi}_{reg}$, the parameter β could be estimated via maximum likelihood or m-estimation. In the next result we give an analog of Theorem 3.1 for the parametric weighting estimator. Note that this setup can be viewed as a semiparametric model in which the propensity score π is restricted to follow a known parametric form, but the regression functions μ_a are left unrestricted.

Theorem 4.3. *Let $f(z) = \left\{ \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right\} y$ and $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$, so that $\psi = \mathbb{E}\{f(Z)\}$ is the average treatment effect. Assume the parametric model*

$$\pi(x) = \pi(x; \beta)$$

for some $\beta \in \mathbb{R}^p$. Suppose the estimator $\widehat{\beta} \in \mathbb{R}^p$ solves an estimating equation so that

$$\mathbb{P}_n\{m(Z; \widehat{\beta})\} = 0.$$

Assume $m(z;\beta) \in \mathbb{R}^p$ is Lipschitz in β, and that $\mathbb{E}\{m(z;\beta)\}$ is differentiable at the true β satisfying $\mathbb{E}\{m(Z;\beta)\} = 0$ with nonsingular derivative matrix. Then

$$\widehat{\psi} - \psi = (\mathbb{P}_n - \mathbb{P}) g(Z; \beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z;\beta) = f(x;\beta) + \frac{\partial \mathbb{E}\{f(X;\beta)\}}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(z;\beta)$$

and so is root-n consistent and asymptotically normal.

Proof. By Lemma 3.1 we have

$$\widehat{\psi} - \psi = Z^* + T_1 + T_2$$

where $Z^* = (\mathbb{P}_n - \mathbb{P})f$ and T_1 and T_2 defined accordingly. By Lemma 3.2 we have

$$\widehat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[\left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n})$$

which also is enough to imply $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$. Further by the delta method we have

$$\begin{aligned} T_2 &= \mathbb{P}(\widehat{f} - f) = h(\widehat{\beta}) - h(\beta) \\ &= (\mathbb{P}_n - \mathbb{P}) \left[\frac{\partial h(\beta)}{\partial \beta^{\mathrm{T}}} \left\{ \frac{\partial \mathbb{E}(m(Z;\beta))}{\partial \beta^{\mathrm{T}}} \right\}^{-1} m(Z;\beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}) \end{aligned}$$

for $h(\beta) = \mathbb{E}\{f(Z;\beta)\} = \mathbb{E}\left[\left\{ \frac{A}{\pi(X;\beta)} - \frac{1-A}{1-\pi(X;\beta)} \right\} Y\right]$. Combining terms gives the result. □

The analysis and interpretation of the weighting estimator parallels that of the regression estimator. In particular, Theorem 4.3 shows that if a correct parametric model is available for the propensity scores π, and under some regularity conditions (essentially smoothness of the model), then the resulting inverse-probability-weighted estimator is root-n consistent and asymptotically normal. Confidence intervals can be constructed using an estimate of the closed-form asymptotic variance $\widehat{\text{var}}\{\widehat{g}(Z;\widehat{\beta})\}$, or via the bootstrap. Surprisingly, another variance estimator is available for the weighting estimator: namely, valid but conservative inference can be obtained via the variance estimate $\widehat{\text{var}}\{f(Z;\widehat{\beta})\}$, which is the variance estimate that would be used if the propensity scores were known to be equal to $\widehat{\pi}(x) = \pi(x;\widehat{\beta})$ [Tsiatis, 2006]

As with regression estimoatrs, the analysis for weighting estimators also needs to be amended in the nonparametric case. Here, the term T_2 can essentially inherit the convergence rate of $\widehat{\pi}$ in the sense that

$$\begin{aligned} \mathbb{P}(\widehat{f} - f) &= \mathbb{P}\left\{ \left(\frac{\pi - \widehat{\pi}}{\widehat{\pi}} \right) \mu_1 - \left(\frac{\widehat{\pi} - \pi}{1 - \widehat{\pi}} \right) \mu_0 \right\} \\ &= \mathbb{P}\left\{ (\pi - \widehat{\pi}) \left(\frac{\mu_1}{\widehat{\pi}} + \frac{\mu_0}{1 - \widehat{\pi}} \right) \right\} \\ &\lesssim \mathbb{P}|\widehat{\pi} - \pi| \leq \sqrt{\mathbb{P}\{(\widehat{\pi} - \pi)^2\}} = \|\widehat{\pi} - \pi\| \end{aligned}$$

where the second inequality follows by Cauchy-Schwarz. Thus we should again not expect root-n rates for nonparametric weighting estimators, and confidence intervals (even using bootstrap) may not be correctly centered, and thus invalid.

However, again similar to the regression case, Hirano et al. [2003] show that undersmoothing can correct these issues, under some assumptions. In particular, they show that under some relatively strong smoothness conditions on the propensity score π (i.e., that π has seven times as many derivatives as it does dimensions d), and undersmoothing, a series-based nonparametric inverse-probability-weighted estimator is root-n consistent and asymptotically normal.

Appendix A

Notation Guide

Y^a	Potential outcome under treatment/exposure $A = a$
$\perp\!\!\!\perp$	Statistically independent
$\xrightarrow{p}$	Convergence in probability
$\rightsquigarrow$	Convergence in distribution
$O_{\mathbb{P}}(1)$	Bounded in probability
$o_{\mathbb{P}}(1)$	Converging in probability to zero
$\mathbb{P}_n$	Sample average operator, as in $\mathbb{P}_n(\widehat{f}) = \mathbb{P}_n\{\widehat{f}(Z)\} = \frac{1}{n}\sum_{i=1}^n \widehat{f}(Z_i)$
$\mathbb{P}$	Conditional expectation given the sample operator, as in $\mathbb{P}(\widehat{f}) = \int \widehat{f}(z)\ d\mathbb{P}(z)$
$\lVert \cdot \rVert$	$L_2(\mathbb{P})$ norm $\lVert f \rVert = \sqrt{\mathbb{P}(f^2)}$ or Euclidean norm, depending on context
$\lVert \cdot \rVert_1$	$L_1(\mathbb{P})$ norm $\lVert f \rVert_1 = \mathbb{P}(\lvert f \rvert)$
$\lVert \cdot \rVert_\infty$	L_∞ or supremum norm $\lVert f \rVert_\infty = \sup_z \lvert f(z) \rvert$
$\mathcal{H}(s)$	Hölder class of functions with smoothness index s
$\lesssim$	Less than or equal, up to a constant multiplier

Bibliography

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.

M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.

S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.

E. H. Kennedy, S. Balakrishnan, and M. G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.

S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.

D. Michaels. *Doubt is their product: how industry's assault on science threatens your health*. Oxford University Press, 2008.

J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.

J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1): 113–124, 2000.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.

Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.